

THEORETICAL ANALYSIS OF DYNAMIC CLUSTERING

BY XIAOGANG WANG

A theoretical framework is derived to investigate the convergence and stability of dynamic clustering methods which transform data according to different laws of attraction to achieve autonomous partitions. On applying the conservation law, we establish partial differential equations to prescribe the successive transformations of the underlying probability densities in dynamic clustering. These partial differential equations correspond to anti-diffusion processes and are solved analytically. We show that a broad class of unsupervised shrinking or clustering methods including the mean-shift algorithm are intrinsically unstable except for independent normal densities. Theoretical results of the supervised dynamic clustering processes indicate that an effective supervision must be chosen judiciously to ensure a correct convergence since a universally optimal supervising function does not exist.

1. Introduction. Clustering is the process of partitioning a set of objects into subgroups according to certain measure of similarity. Cluster analysis has many applications in data mining in which large data sets, such as biological data or climate data, need to be partitioned into much smaller and homogeneous groups. Hastie et al. (2001) and Han and Kamber (2006) both offer excellent reviews on clustering algorithms with different emphasis.

For many classical clustering algorithms, such as K-means (MacQueen 1967; Hartigan and Wong 1979) and PAM (Kaufman and Rousseeuw 1990), the number of clusters or sub-populations needs to be specified by the user.

AMS 2000 subject classifications: Primary 62H20

Keywords and phrases: anti-diffusion, convergence, dynamic clustering, , partial differential equations

One approach is to select the number of clusters by optimizing a certain measure of strength of the clusters (Tibshirani, Walther and Hastie 2000 and Fraley and Raftery 2002). An alternative is to first partition the data into many small clusters, and then merge these small clusters until no clusters can be merged (Frigui and Krishnapuram 1999). The third strategy is to extract one cluster at a time (Zhung et al. 1996). The determination of the number of clusters, however, remains to be a challenging problem when the clusters assume non-normal shapes with blurring or even slightly overlapping boundaries. It can also be very difficult to specify the exact functional form of the underlying probability density function due to the uncertainty involved in the clustering processes.

Many automatic or non-parametric clustering algorithms have emerged from various disciplines in the last twenty years with a wide range of applications to pattern recognition and image analysis. Although there are some minor technical or operational differences, they all treat data points as autonomous agents or particles and iteratively move them toward cluster centers or focal points. One approach, the gravitational clustering (Wright 1977; Kundu 1999; Sato 2000; Wang and Rau 2001), considers each data point as a particle of unit mass with zero velocity which is gradually moving toward the cluster center by imposing the gravitational law. The theoretical properties of gravitational clustering, however, have not been fully established although the original idea has been proposed for thirty years.

At the mean time, non-parametric methods have received increasing attentions in the literature due to its flexible and adaptive nature. The most famous and representative method is the so-called *mean-shift* algorithm proposed by Fukunaga and Hostetler (1975), Cheng (1995), Comaniciu and Meer,(2002). Given a kernel function K and a weight function w , the gen-

eralized mean-shift operation is given by

$$(1.1) \quad T(x) = \frac{\sum K(x, s) w(s) s}{\sum K(x, s) w(s)}.$$

It originates from the ideas in following the gradient in kernel density estimation since data points are transformed toward denser regions by using a functional of the kernel functions. There are many variations of this algorithm. For example, Virmajoki (2002) Shi et al. (2005) and Wang et al. (2007a) propose algorithms that closely resemble the main idea of the mean-shift method. Although data sharpening procedure proposed by Choi and Hall (1999) is originally designed to reduce bias by pushing data points at the boundary a bit closer to the center, the movements also resemble the one in the mean-shift method. Woolfold and Braun (2006) apply the data sharpening method for the identification and tracking of spatial temporal centers of lightning activity.

Although non-parametric clustering methods are appealing to practitioner and have been applied in many research areas, the underlying process is not well understood since the initial probability density function undergoes continuous nonlinear transformations. Chen (1995) pointed out that it is difficult to see where the mean-shift method leads to since all data points are moving simultaneously. The same statement is also true for almost all non-parametric dynamic clustering methods. Although local properties of these algorithms are intuitively clear, the intrinsic global patterns have proven to be difficult to establish. The major difficulty to gain insights about the validity of these methods is due to complex spatial and temporal patterns embedded within these dynamic clustering methods.

In this article, we develop an analytical framework for dynamic clustering which prescribes the time space evolution of the initial probability density

function. The mathematical properties derived from the framework provide guidance and determinism for complex dynamic evolutions with clear criteria to evaluate convergence and reliability for a general class of dynamic clustering methods. The proposed framework is derived by modeling the successive nonlinear transformations of the underlying probability density function based the general conservation law which describes the conservation of some basic physical quantity of a closed system. The dynamic clustering process is then characterized by studying the time-space evolution of the initial probability density function. The differential form of the conservation law derived from the dynamic clustering scheme is a class of second order partial differential equations (PDE) that are anti-diffusive in nature. On applying the proposed framework, we find that a broad class of unsupervised dynamic clustering methods only converges for normal densities with independent structures. The entropy of the corresponding clustering process has been proven to be non-increasing through time. This is a direct violation of the second law in thermodynamics. Consequently, these clustering processes are unnatural and could exhibit chaotic behaviors locally at any given time. Therefore, supervised clustering should be preferred in order to ensure a general and correct convergence. Theoretical results for supervised dynamic clustering indicate that a universally effective supervising function does not exist. Therefore, each supervising function should be chosen judiciously for each specific application.

This article is organized as follows. We present the general framework for dynamic clustering in Section 2. Section 3 presents theoretical analysis of unsupervised dynamic clustering. The convergence for supervised dynamic clustering is shown in Section 4. Discussions are provided in Section 5.

2. Theoretical Framework for Dynamic Clustering. We establish a theoretical framework that prescribes recurrent regularities during the shrinking or clustering processes.

2.1. Evolutions of Densities. In dynamic clustering, each data point is considered as a particle under a gravitational field or an autonomous agent governed by certain laws of attractions. Although the realization of these clustering methods is discrete, one can imagine an underlying continuous process when the total number of iterations is very large. In such a continuous process, each data point then travel continuously toward a local target at any given time. Therefore, the underlying probability density function undergoes a constant transformation process. Consequently, this kind of continuous process can be modeled by describing the patterns in the evolution of densities.

We now define a family of probability density function

$$(2.1) \quad \mathcal{E} = \{f_t \mid f(\mathbf{x}; t) \geq 0, \int f(\mathbf{x}; t) d\mathbf{x} = 1, t \in N, \mathbf{x} \in R^n\}.$$

Therefore the space-time evolution of the dynamic clustering processes can then be modeled using probability densities that are also functions of time. Instead of modeling the individual movements of one particular data point, one can examine the pattern of transformations of underlying probability distribution. We now an analytical framework based on some fundamental principles of these processes.

2.2. One-Dimensional Conservation Law . We note that the total number of data points remain the same despite of usage of different law of attractions. Even if some merging strategy is employed, a data point should not lose its deserved influence to the partition process to avoid biased or

inaccurate estimate of the underlying probability density function. The conservation law is perfectly applicable since the rate of change of the total number of particles contained in a fixed volume is equal to the influx of particles or data points passing through the boundary.

To illustrate, we look at the one dimensional case. Denote the one dimensional influx of data points by $q(x; t)$ and the probability density at time t by $f(x; t)$. We then have

$$(2.2) \quad q(x; t) = u(x; t) \times f(x; t),$$

where $u(x; t)$ is the speed of particles at location x and time t .

We can imagine a constant flow of data points passing through an arbitrary small interval due to the utilized laws of attractions. The data points are assumed to be incompressible with constant velocity in this small region. Using the standard argument in fluid dynamics, the one-dimensional *conservation law* is then given by

$$(2.3) \quad \frac{d f(x; t)}{dt} + \frac{d q(x; t)}{dx} = 0.$$

It characterizes the functional connection between the probability density function and the influx function of data points at a given location and time. This provides the fundamental component to establish an analytical framework to describe the space-time evolution of the probability density function.

2.3. The General Partial Differential Equations. We now present the general differential form to incorporate the spatial dimensions. Denote an influx vector by $\mathbf{q}(\mathbf{x}; t)$ and probability density function $f(\mathbf{x}; t)$. We then have

$$(2.4) \quad \frac{d}{dt} \int_V f(\mathbf{x}; t) dV = - \int_S (\mathbf{q} \cdot \mathbf{n}) dS,$$

where dV is the volume element and dS is the surface element of the boundary surface S , \mathbf{n} denotes the outward unit normal vector to S and the right-hand side measures the *outward* influx indicated by the minus sign.

On applying the Gauss divergence theorem and taking d/dt inside of the integral on the left hand side, we then have

$$(2.5) \quad \int_V \left(\frac{\partial f(\mathbf{x}, t)}{\partial t} + \nabla \cdot \mathbf{q}(\mathbf{x}, t) \right) dV = 0.$$

where ∇ is the divergence operator. given by

$$(2.6) \quad \nabla \cdot \mathbf{q}(\mathbf{x}, t) = \sum_{i=1}^n \frac{\partial q_i(\mathbf{x}, t)}{\partial x_i},$$

where q_i 's are the coordinate functions of $\mathbf{q}(\mathbf{x}, t)$.

Since the result is valid for any arbitrary volume V , the integrand must be zero if it is continuous. The differential form of the general conservation law is then given by

$$(2.7) \quad \frac{\partial f(\mathbf{x}, t)}{\partial t} + \nabla \cdot \mathbf{q}(\mathbf{x}, t) = 0,$$

where $\mathbf{q}(\mathbf{x}; t) = \mathbf{u}(\mathbf{x}; t) \times \mathbf{f}(\mathbf{x}; t)$. Detailed derivations and discussions of conservation laws and associated differential forms can be found in Debneth (2004).

For supervised dynamic clustering, the trajectories of data points will also be influenced or dictated by supervision. This is equivalent of imposing a source or sink function in the process since data points will be “absorbed” in a given domain. Denote the supervision function or sink function by $\psi(\mathbf{x}, t)$. Following a similar argument, the conservation law with a sink function is given by

$$(2.8) \quad \frac{\partial f(\mathbf{x}, t)}{\partial t} + \nabla \cdot \mathbf{q}(\mathbf{x}, t) = \psi(\mathbf{x}, t).$$

This framework is applicable to many dynamic clustering processes as long as data points are not lost in the partition process. It provides the much needed an analytical framework and established the foundation of our subsequent theoretical analysis of dynamic clustering methods. We will establish theoretical results for unsupervised dynamic clustering methods including the well known mean-shift method in the next section.

3. Properties of Unsupervised Dynamic Clustering .

3.1. *Unsupervised Dynamic Clustering.* In unsupervised dynamic clustering methods, the movements of data points depend on the functional connection between the current probability density function and its gradient or first order derivative. The movements of many non-parametric dynamic clustering methods are often governed by a law such that a data point will move to the center along more or less the direction of the gradient adjusted by the value of the current density function at the point of interest.

This can be formulated mathematically as the following:

$$(3.1) \quad u(\mathbf{x}; t) = a^2 \frac{\nabla f(\mathbf{x}; t)}{f(\mathbf{x}; t)}$$

The gradient component forces each data point to optimize its trajectory to seek a local mode or cluster center which is known as the *mode seeking* property. The movement is also proportional to the reciprocal value of the current probability density function. This implies that data points in sparsely populated areas will travel longer distances when compared with that of data points in densely populated area even if the gradient functions assume the same value at these two different locations.

Following the argument in Cheng(1995), one can show that the mean-shift algorithm indeed belongs to this category. All the different variations

or improved versions based on the mean-shift method therefore are embraced by this category as well. As an alternative to the traditional movements in the mean-shift method, data points could move to the local center given by the conditional mean:

$$(3.2) \quad \mathbf{x}^{k+1} = \frac{\int_{B(\mathbf{x}^k, d)} \mathbf{t} f(\mathbf{t}) d\mathbf{t}}{\int_{B(\mathbf{x}^k, d)} f(\mathbf{t}) d\mathbf{t}},$$

where $B(\mathbf{x}^k, d)$ is a neighborhood with the center located at \mathbf{x}^k and the radius d . Wang et al. (2007b) showed that

$$(3.3) \quad \mathbf{x}^{k+1} = \mathbf{x}^k + \frac{n}{n+2} \frac{d^2}{f(\mathbf{x}; t)} \nabla f(\mathbf{x}; t) + O(d^3).$$

Since a K -nearest neighbor approach provides a natural estimate for a conditional mean, it will also satisfy equation (3.1).

3.1.1. *Convergence Analysis for One-Dimensional Case.* By virtue of eqn (2.4) and the assumption by eqn(3.1), the corresponding differential form is then given by

$$(3.4) \quad \frac{\partial}{\partial t} f(x; t) = -a^2 \frac{d^2 f(x; t)}{dx^2},$$

where $a > 0$ and $f(x, 0) = \phi_0(x)$, the initial probability density function.

This is a one-dimensional anti-diffusion equation. Anti-diffusion processes are rare in reality and they do not appear often in the literature except for some special types such as the crystallization process. Therefore, anti-diffusion equations have quite unique characteristics than those from diffusion or other equations. We now present the exact analytical solution to this differential equation.

THEOREM 3.1. *Under assumption by equation (3.1), the one-dimensional*

anti-diffusion equation has one unique solution and takes the following form

$$(3.5) \quad f(x; t) = \frac{1}{\sqrt{-4a^2\pi t}} \int_{-\infty}^{\infty} \phi_0(\xi) e^{-\frac{(\xi-x)^2}{-4a^2t}} d\xi, \quad t \leq 0,$$

where $f_0(x) = \phi_0(x)$, the initial probability density function.

Proof: Consider the Fourier transformation of $f(x; t)$:

$$F(\omega; t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x; t) e^{i\omega x} dx; \quad f(x; t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega, t) e^{-i\omega x} d\omega.$$

First, observe that

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{\partial}{\partial t} f(x; t) e^{i\omega x} dx &= \frac{\partial}{\partial t} \left(\int_{-\infty}^{\infty} f(x; t) e^{i\omega x} dx \right) \\ &= \frac{\partial}{\partial t} F(\omega; t). \end{aligned}$$

Furthermore, note that

$$\begin{aligned} \int_{-\infty}^{\infty} f_{xx}(x; t) e^{i\omega x} dx &= \int_{-\infty}^{\infty} e^{i\omega x} \frac{d}{dx} (f_x(x; t)) \\ &= e^{i\omega x} f_x(x; t) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{i\omega x} f_x(x; t) dx \\ &= \int_{-\infty}^{\infty} f(x; t) e^{i\omega x} dx \\ &= -\omega^2 F(\omega, t). \end{aligned}$$

It then follows from equation (3.4) that

$$(3.6) \quad \frac{\partial}{\partial t} F(\omega; t) - a^2 \omega^2 F(\omega, t) = 0.$$

The initial boundary condition also gives rise to

$$(3.7) \quad \Phi_0(\omega) = F(\omega; 0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi_0(x) e^{i\omega x} dx.$$

Consequently, the equation (3.6) has the solution

$$(3.8) \quad F(\omega; t) = \Phi_0(\omega) e^{a^2 \omega^2 t}.$$

It then follows that

$$(3.9) \quad f(x; t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Phi_0(\omega) e^{a^2 \omega^2 t - i\omega x} d\omega.$$

Note that the domain of convergence for the above integral is $(-\infty, 0)$.

We then have

$$\begin{aligned} f(x; t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \phi_0(\xi) e^{i\omega \xi} d\xi \right) e^{a^2 \omega^2 t - i\omega x} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_0(\xi) \left(\int_{-\infty}^{\infty} e^{a^2 \omega^2 t + i\omega(\xi - x)} d\omega \right) d\xi. \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_0(\xi) \left(\sqrt{\frac{\pi}{a^2(-t)}} e^{(\xi - x)^2 / 4a^2 t} \right) d\xi. \\ &= \frac{1}{\sqrt{-4a^2 \pi t}} \int_{-\infty}^{\infty} \phi_0(\xi) e^{-\frac{(\xi - x)^2}{-4a^2 t}} d\xi, \quad t \leq 0. \quad \diamond \end{aligned}$$

The fact that the solution is only available for $t \leq 0$ implies that the evolutions of densities are deterministic causal events. Given current status, there is only one unique process that will cause or explain what has occurred. This is proven to be a powerful tool to establish convergence analysis for dynamic clustering methods as shown in the next theorem.

THEOREM 3.2. *Under assumption by equation (3.1), a convergence to a location μ_0 can only occur for normal densities with mean μ_0 and variance proportional to a^2 .*

In addition, the first order derivative of the variance with respect to time is given by

$$(3.10) \quad \frac{d\sigma_t^2}{dt} = -2a^2.$$

where σ_t^2 denotes the variance of the normal density at time t .

The converging speed of a data point at a location x at time t is

$$(3.11) \quad u(x; t) = \frac{x - \mu_0}{2a^2(-t)}, \quad t \leq 0.$$

Proof: If the convergence at time $t = 0$ at a location μ_0 , this implies that $f_0 = \delta(x - \mu_0)$. By Theorem 3.1, it then follows that

$$(3.12) \quad f(x; t) = \frac{1}{\sqrt{-4a^2\pi t}} e^{-\frac{(x-\mu_0)^2}{-4a^2t}}, \quad t \leq 0.$$

The variance takes the form $2a^2(-t)$. Therefore, $\frac{d\sigma_t^2}{dt} = -2a^2$. The rest of the result follow immediately from the assumption. \diamond

The assertion from this theorem effectively states that a convergence can only occur for a normal density for one dimensional case. The spatial variation of normal densities at different time point depends on the magnitude of contraction forced by the attraction law. We show that this result also holds true for higher-dimensional spaces when there are multiple cluster centers.

3.1.2. *Convergence in Multi-dimensional Space.* Under assumption by equation (3.1), it then follows that

$$(3.13) \quad \mathbf{q}(\mathbf{x}, t) = u(\mathbf{x}, t) * f(\mathbf{x}, t) = a^2 \nabla f(\mathbf{x}, t).$$

Consequently, the equation (2.7) now becomes

$$(3.14) \quad \frac{\partial f(\mathbf{x}, t)}{\partial t} + a^2 \nabla^2 f(\mathbf{x}, t) = 0,$$

where $\text{the Laplacian of } f \nabla^2 f = \sum_{i=1}^n \partial^2 f / \partial x_i^2$, and the boundary condition is given by $f(\mathbf{x}, 0) = f_0(\mathbf{x})$.

The phenomenon for the one dimensional case be generalized to the multi-dimensional case when there are multiple cluster centers.

THEOREM 3.3. Under the assumption specified in equation (3.1), the anti-diffusion equation

$$(3.15) \quad \frac{\partial f(\mathbf{x}, t)}{\partial t} + a^2 \nabla^2 f(\mathbf{x}, t) = 0,$$

with the boundary condition $f|_{t=0} = \phi_0$ has the solution

$$(3.16) \quad f(\mathbf{x}, t) = (-4a^2 t)^{-n/2} \int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta}) e^{-\frac{(\boldsymbol{\eta}-\mathbf{x})^2}{-4a^2 t}} d\boldsymbol{\eta},$$

where $(\boldsymbol{\eta} - \mathbf{x})^2 = \sum_{i=1}^n (\eta_i - x_i)^2$.

Proof: The dynamic shrinking or clustering can be characterized as

$$(3.17) \quad \frac{\partial f(\mathbf{x}, t)}{\partial t} + a^2 \nabla^2 f(\mathbf{x}, t) = 0,$$

with boundary condition $f(\mathbf{x}, 0) = \phi_0(\mathbf{x})$, a probability density function.

Denote the n -dimensional Fourier transformation as the following:

$$(3.18) \quad F_n(\mathbf{s}; t) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\mathbf{x}, t) e^{i \mathbf{s} \cdot \mathbf{x}} dx_1 dx_2 \dots dx_n,$$

where $\mathbf{s} \cdot \mathbf{x} = \sum_{i=1}^n s_i x_i$.

Apply Fourier transformation on both sides of equation (3.17), we then have

$$(3.19) \quad \partial F_n(\mathbf{s}, t) / \partial t + a^2 \|\mathbf{s}\| F_n(\mathbf{s}, t) = 0$$

where $\|\mathbf{s}\| = \sum_{i=1}^n s_i^2$.

Thus, the solution for above equation is

$$(3.20) \quad F_n(\mathbf{s}, t) = F_n(\mathbf{s}, 0) e^{-a^2 \|\mathbf{s}\| t}.$$

where

$$(3.21) \quad F_n(\mathbf{s}, 0) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \int_{(\mathbf{x})} \phi_0(\mathbf{x}) e^{i \mathbf{s} \cdot \mathbf{x}} d\mathbf{x}.$$

Using inverse Fourier transformation, we then have

$$(3.22) \quad f(\mathbf{x}, t) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \int_{(\mathbf{s})} F_n(\mathbf{s}, 0) e^{-a^2 \|\mathbf{s}\|^2 t - i \mathbf{s} \cdot \mathbf{x}} d\mathbf{s}.$$

It then follows that

$$(3.23) \quad f(\mathbf{x}, t) = \left(\frac{1}{2\pi} \right)^n \int_{(\mathbf{s})} \left(\int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta}) e^{i \mathbf{s} \cdot \boldsymbol{\eta}} d\boldsymbol{\eta} \right) e^{-a^2 \|\mathbf{s}\|^2 t - i \mathbf{s} \cdot \mathbf{x}} d\mathbf{s}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)$.

By rearranging the order of integration and simplifying the same way in Theorem 3.1, we then have

$$(3.24) \quad f(\mathbf{x}, t) = (-4a^2 t \pi)^{-n/2} \int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta}) e^{-\frac{(\boldsymbol{\eta} - \mathbf{x})^2}{-4a^2 t}} d\boldsymbol{\eta},$$

where $(\boldsymbol{\eta} - \mathbf{x})^2 = \sum_{i=1}^n (\eta_i - x_i)^2$. \diamond

The analytic solution allows us to retract the stages of the convergence and recover the deterministic pattern of evolution of the initial probability density function. We show that the family of probability density functions that guarantee a convergence is a multivariate normal distribution with independent correlation structure.

THEOREM 3.4. *Under the assumption by equation (3.1), a dynamic shrinking or clustering converges to m distinct cluster centers if and only if the initial density function is a mixture of normal distribution with equal variances, i.e.*

$$(3.25) \quad f(\mathbf{x}, t) = \sum_{i=1}^m \lambda_i \phi_i, \quad t < 0.$$

where ϕ_i is the normal density function with mean $\boldsymbol{\mu}_i$ and variance $-2a^2 t$.

Proof:

If the dynamic process converges to a finite number of focal points, *i.e.*,

$$(3.26) \quad \phi_0(\boldsymbol{\eta}) = \sum_{j=1}^m \lambda_j \delta(\boldsymbol{\eta} - \boldsymbol{\mu}_j), \quad \lambda_j \geq 0 \quad \text{and} \quad \sum_{j=1}^m \lambda_j = 1,$$

where $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2}, \dots, \mu_{jn})$, and $\delta(\boldsymbol{\eta} - \boldsymbol{\mu}_j) = \prod_{i=1}^n \delta(\eta_j - \mu_{ji})$.

It then follows that

$$(3.27) \quad f(\boldsymbol{x}, t) = \sum_{j=1}^m \lambda_j \left(\frac{1}{2\pi\sigma_t^2} \right)^{-n/2} e^{-\frac{(\boldsymbol{x} - \boldsymbol{\mu}_j)^2}{2\sigma_t^2}},$$

where $\sigma_t^2 = -2a^2t$, $t < 0$. \diamond

This theorem also implies that contraction rates of dynamic shrinking or clustering are homogenous in all directions for a correct convergence. An heterogeneous contraction pattern therefore implies a non-convergence.

3.2. *Instability of Unsupervised Dynamic Shrinking.* Although this general class of dynamic clustering method seem to be intuitively appealing and a practical convergence might be observed, our theoretical analysis has revealed that they actually will not converge correctly except for normal densities with independent structure. Therefore an observed *convergence* using unsupervised dynamic clustering for non-normal densities or normal densities with non-trivial correlation structures is then either an illusion or an artificial and incorrect one. For clusters that are well separated in low dimensions, a correct convergence could occur if the smoothing parameter is set to be suitably large to capture local patterns accurately. However, there is little hope to expect this kind of success for high-dimensional data when the manifolds are complex and not well separated.

In order to understand a little more precisely a possible instability, we examine the variations of the system through time using the quantity called entropy. The entropy has been widely used in information theory and coding theory. The entropy for a given density function f is defined by

$$(3.28) \quad H(f) = - \int f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x},$$

The entropy is a measure of uncertainty. For example, the entropy for a normal distribution with mean μ and variance σ^2 is $\frac{1}{2} \log(2\pi\sigma^2 + 1)$.

We shall not be concerned with the information-theoretic aspects of the entropy in this article. Instead we will use its original utility to describe the variability of a dynamic process. We now prove the following theorem which prescribes the change of entropy.

THEOREM 3.5. *Assume that the conservation law assumption described in equation (3.1) hold. If*

$$(3.29) \quad \lim_{x_i \rightarrow \infty} \log f(\mathbf{x}) \frac{\partial f(\mathbf{x})}{\partial x_i} = 0, \quad i = 1, 2, \dots, m,$$

then

$$(3.30) \quad \frac{dH(f_t)}{dt} < 0.$$

Proof: Consider the first order derivative of $H(f_t)$ with respect to t . We then have

$$\begin{aligned} \frac{dH(f_t)}{dt} &= - \int_{-\infty}^{\infty} \frac{\partial}{\partial t} (f(\mathbf{x}) \log f(\mathbf{x})) d\mathbf{x} \\ &= - \int_{-\infty}^{\infty} \left(\frac{df(\mathbf{x})}{dt} \log f(\mathbf{x}) + \frac{1}{f(\mathbf{x})} f(\mathbf{x}) \frac{df(\mathbf{x})}{dt} \right) d\mathbf{x} \\ &= - \int_{-\infty}^{\infty} (1 + \log f(\mathbf{x})) \frac{df(\mathbf{x})}{dt} d\mathbf{x} \\ &= \int_{-\infty}^{\infty} (1 + \log f(\mathbf{x})) a^2 (\nabla^2 f(\mathbf{x}, t)) d\mathbf{x} \\ &= \int_{-\infty}^{\infty} (1 + \log f(\mathbf{x})) a^2 \left(\sum_{i=1}^n \frac{\partial^2 f(\mathbf{x}, t)}{\partial x_i^2} \right) d\mathbf{x} \\ &= - \int_{-\infty}^{\infty} a^2 \frac{1}{f(\mathbf{x})} \sum_{i=1}^n \left(\frac{\partial f(\mathbf{x})}{\partial x_i} \right)^2 d\mathbf{x} < 0. \end{aligned}$$

The last step follows from the integration by part and the assumption of the theorem. \diamond .

This shows that the entropy is non-increasing at all times and therefore dynamic clustering method incur a complete violation of the second law of thermodynamics. So the dynamic shrinking and clustering do not correspond to a natural process and are unstable in nature except for normal densities. Without the intervention of supervision, these artificial laws defined locally could produce chaotic and unpredictable movements.

4. Convergence with Supervision. We have established that the convergence of unsupervised shrinking or clustering can only be achieved for independent normal random variables with equal variances. A natural question is whether this instability or non-convergence can be overcome by

some kind of supervision. Mathematically, this might be equivalent to imposing a so called *sink* function into the PDE in the following form:

$$(4.1) \quad \frac{\partial f(\mathbf{x}, t)}{\partial t} + a^2 \nabla^2 f(\mathbf{x}, t) = \psi(\mathbf{x}, t),$$

where ψ is a continuous function.

THEOREM 4.1. *Under the assumption by equation (3.1), the PDE associated with supervised clustering has the following solution*

$$(4.2) \quad \begin{aligned} f(\mathbf{x}, t) = & (-4a^2 t \pi)^{-n/2} \int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta}) e^{-\frac{(\boldsymbol{\eta}-\mathbf{x})^2}{-4a^2 t}} d\boldsymbol{\eta} \\ & + \int_t^0 \int_{(\boldsymbol{\xi})} \psi(\boldsymbol{\xi}, \tau) [-4a^2(t-\tau)\pi]^{-n/2} e^{-\frac{(\mathbf{x}-\boldsymbol{\eta})^2}{-4a^2(t-\tau)}} d\boldsymbol{\xi} d\tau, \quad t \leq 0. \end{aligned}$$

If the clustering process converges to m distinct focal points, then

$$(4.3) \quad \begin{aligned} f(\mathbf{x}, t) = & \sum_{j=1}^m \lambda_j \left(\frac{1}{2\pi\sigma_t^2} \right)^{-n/2} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu}_j)^2}{2\sigma_t^2}} \\ & + \int_t^0 \int_{(\boldsymbol{\xi})} \psi(\boldsymbol{\xi}, \tau) [-4a^2(t-\tau)\pi]^{-n/2} e^{-\frac{(\mathbf{x}-\boldsymbol{\eta})^2}{-4a^2(t-\tau)}} d\boldsymbol{\xi} d\tau, \quad t \leq 0. \end{aligned}$$

Proof:

The general solution with the sink function in the PDE can be decomposed into two parts:

$$(4.4) \quad f(\mathbf{x}, t) = g_1(\mathbf{x}, t) + g_2(\mathbf{x}, t),$$

where g_1 is the solution for the PDE in eqn(3.15) with boundary condition $g_1(t=0) = \phi_0$ and g_2 satisfying the PDE specified by eqn(4.1) with the boundary condition $g_{t=0}^2 = 0$.

(ii) The function form of g_1 is given by Theorem 3. That is,

$$g_1(\mathbf{x}, t) = (-4a^2 t \pi)^{-n/2} \int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta}) e^{-\frac{(\boldsymbol{\eta}-\mathbf{x})^2}{-4a^2 t}} d\boldsymbol{\eta},$$

where $(\boldsymbol{\eta} - \mathbf{x})^2 = \sum_{i=1}^n (\eta_i - x_i)^2$.

(iii) To find g_2 , we consider a nonhomogeneous differential equation of the form

$$(4.5) \quad L_{\mathbf{x}} u(\mathbf{x}) = \psi(\mathbf{x}, t),$$

where $L_{\mathbf{x}}$ is a linear partial differential operator associated with the initial PDE, that is

$$(4.6) \quad L_{\mathbf{x}} u(\mathbf{x}) = \frac{\partial u(\mathbf{x}, t)}{\partial t} + a^2 \nabla^2 u(\mathbf{x}, t).$$

The Green function $G(\mathbf{x}, \boldsymbol{\xi})$ of this problem satisfies the equation

$$(4.7) \quad L_{\mathbf{x}} G(\mathbf{x}, \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}) \delta(t - \tau)$$

where $G_{t=0} = 0$.

The solution for the partial differential equation by eqn(4.5) is then given by

$$(4.8) \quad u(\mathbf{x}) = \int_t^0 \int_{(\boldsymbol{\xi})} \psi(\boldsymbol{\xi}, \tau) G(\mathbf{x}, t; \boldsymbol{\xi}, \tau) d\boldsymbol{\xi} d\tau,$$

where the Green function satisfying the following PDE

$$\begin{aligned} \frac{\partial G(\mathbf{x}, t)}{\partial t} + a^2 \nabla^2 G(\mathbf{x}, t) &= 0, \\ G|_{t=\tau} &= \delta(\mathbf{x} - \boldsymbol{\xi}). \end{aligned}$$

By theorem 3 and replacing t by $t - \tau$, the Green function is then given by

$$\begin{aligned} (4.9) \quad G(\mathbf{x}, t) &= [-4a^2(t - \tau)\pi]^{-n/2} \int_{(\boldsymbol{\eta})} \delta(\boldsymbol{\eta} - \boldsymbol{\xi}) e^{-\frac{(\boldsymbol{\eta}-\mathbf{x})^2}{-4a^2(t-\tau)}} d\boldsymbol{\eta} \\ &= [-4a^2(t - \tau)\pi]^{-n/2} e^{-\frac{(\mathbf{x}-\boldsymbol{\eta})^2}{-4a^2(t-\tau)}} \end{aligned}$$

where $(\mathbf{x} - \boldsymbol{\eta})^2 = \sum_{i=1}^n (x_i - \eta_i)^2$.

It then follows that

$$(4.10) \quad g_2(\mathbf{x}, t) = \int_t^0 \int_{(\boldsymbol{\xi})} \psi(\boldsymbol{\xi}, \tau) [-4a^2(t - \tau)\pi]^{-n/2} e^{-\frac{(\mathbf{x} - \boldsymbol{\xi})^2}{-4a^2(t - \tau)}} d\boldsymbol{\xi} d\tau, \quad t \leq 0.$$

Therefore,

$$(4.11) \quad \begin{aligned} f(\mathbf{x}, t) = & \frac{1}{M} (-4a^2 t \pi)^{-n/2} \int_{(\boldsymbol{\eta})} \phi_0(\boldsymbol{\eta}) e^{-\frac{(\mathbf{x} - \boldsymbol{\eta})^2}{-4a^2 t}} d\boldsymbol{\eta} \\ & + \frac{1}{M} \int_t^0 \int_{(\boldsymbol{\xi})} \psi(\boldsymbol{\xi}, \tau) [-4a^2(t - \tau)\pi]^{-n/2} e^{-\frac{(\mathbf{x} - \boldsymbol{\xi})^2}{-4a^2(t - \tau)}} d\boldsymbol{\xi} d\tau, \quad t \leq 0, \end{aligned}$$

where M is the normalizing constant to ensure that $f(\mathbf{x}, t)$ is a proper probability density function. \diamond

The fact that the original density function of the PDE is a functional of the supervision function implies that a correct convergence is dependent on the choice of the supervising function. The assertion of the theorem indicates that a universally effective supervising function does not exist. A supervising function then must be chosen judiciously to ensure a correct convergence.

We remark that there are some dynamic shrinking and clustering algorithms that could be stable due to external sink functions. This is due to the violation of the conservation law. One possible scenario is the crystallization processes as described in Teran and Bill (2010). It is stable due to the fact that particles are accumulating and transformed into solid with zero speed due to the crystallization.

5. Discussion. Convergence and stability analysis of dynamic clustering methods are almost non-existent in the literature. By using the conservation law, we establish partial differential equations that prescribe the spatial and temporal variations and evolutions of these clustering processes. We show that, in the absence of a sink function or intervention, many dynamic clustering methods including the well known mean-shift algorithm do not result in a correct convergence in general unless all variables involved are independent and normally distributed. The non-increasing property of the entropy and the anti-diffusive nature of many dynamic clustering methods prevent them to be universally reliable without a proper supervision. A supervised dynamic clustering should be preferred and the supervising function must be chosen carefully to ensure valid results.

We emphasize that an artificial convergence can be generated by setting certain parameters such as the radius or smoothing parameter to certain values. For example, a global convergence to one focal point can be achieved if the parameters are set such that all data points will be forced to move to one cluster. A practical convergence can also be produced by many suitable values. Although such a convergence can be achieved, this is usually not meaningful since significant local geometrical properties will not be respected in those circumstances.

A semi-dynamic or static clustering method could produce a correct convergence if only one data point is allowed to move while the entire probability density function remains unchanged. The entire process is only dynamic for a chosen data point. This class of method is often not preferred due to its slow convergence.

Work is in progress to extend the framework proposed in this article to provide theoretical analysis for other clustering methods.

References.

- [1] Choi, E. and Hall, P. (1999). Data sharpening as a prelude to density estimation *Biometrika*, **86**, 941-947.
- [2] Cover, T.M. and Thomas, J.A. (2006) *Elements of Information Theory*. John Wiley & Sons, Inc. New Jersey.
- [3] Cheng, L. (1995) Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**, 790-799.
- [4] Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE transactions on pattern analysis and machine intelligence*, 24(5): 603-619.
- [5] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611-631.
- [6] Frigui, H. and Krishnapuram, R. (1999). A robust competitive clustering algorithm with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):450-465.
- [7] Fukunaga, K. and Hostetler, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21:32-40.
- [8] Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*, 2nd edition. The Morgan Kaufmann Series in Data Management Systems.
- [9] Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, 28:100-108.
- [10] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag.
- [11] Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- [12] Kundu, S. (1999). Gravitational clustering: a new approach based on the spatial distribution of the points. *Pattern Recognition*, 32:1149-1160.
- [13] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, pages 281-297. Berkeley, Calif: University of California Press.
- [14] Sato, Y. (2000). An autonomous clustering technique. In Kiers, A. L. Henk, Ras-

- son, Jean-Paul, Groenen, Patrick J. E., and Schader, Martin, editors, *Data analysis, classification, and related methods*. Springer.
- [15] Shi, Y.; Song, Y.; Zhang, A. (2005) A shrinking-based clustering approach for multi-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 1389-1403.
- [16] Teran, A. V. and Bill, A. (2010). Time-evolution of grain size distributions in random nucleation and growth crystallization processes. *Physics Review*, **81**, 19.
- [17] Tibshirani, R., Walther, G., and Hastie, T. (2000). Estimating the number of clusters in a dataset via the gap statistic. *Technical Report 208*, Dept. of Statistics, Stanford University.
- [18] Virmajoki, O.; Franti, P.; Kaukoranta, T. (2002) Iterative shrinking method for generating clustering. *Proceedings of the International Conference on Image Processing*, **2**, 685-688.
- [19] Wang, J. H. and Rau, J. D. (2001). VQ-agglomeration: A novel approach to clustering. *IEE Proceedings-Vision, Image and Signal Processing*, 148(1):36-44.
- [20] Wang, X., Qiu, W. and Zamar, H. R. (2007a). CLUES: A non-parametric clustering method based on local shrinking. *Computational Statistics and Data Analysis*, **52**, 286-298.
- [21] Wang, X., Liang, D, Feng, X. and Ye, L. (2007b) A derivative-free optimization algorithm based on conditional moments, *Mathematical Analysis and Applications*, **331**, 1337-1360.
- [22] Woolfold, D. G. and Braun, W. J. (2006). Convergent data sharpening for the identification and tracking of spatial temporal centers of lightning activity, *Envirometrics*, **18**, 461-479.
- [23] Wright, W. E. (1977). Gravitational clustering. *Pattern Recognition*, 9:151-166.
- [24] Zhung, X., Huang, Y., Palaniappan, K., and Lee, J. S. (1996). Gaussian mixture modeling, decomposition and applications. *IEEE Transactions on Signal Process*, 5: 1293-1302.

XIAOGANG(STEVEN) WANG
DEPARTMENT OF MATHEMATICS AND STATISTICS
YORK UNIVERSITY
CANADA
E-MAIL: stevenw@mathstat.yorku.ca